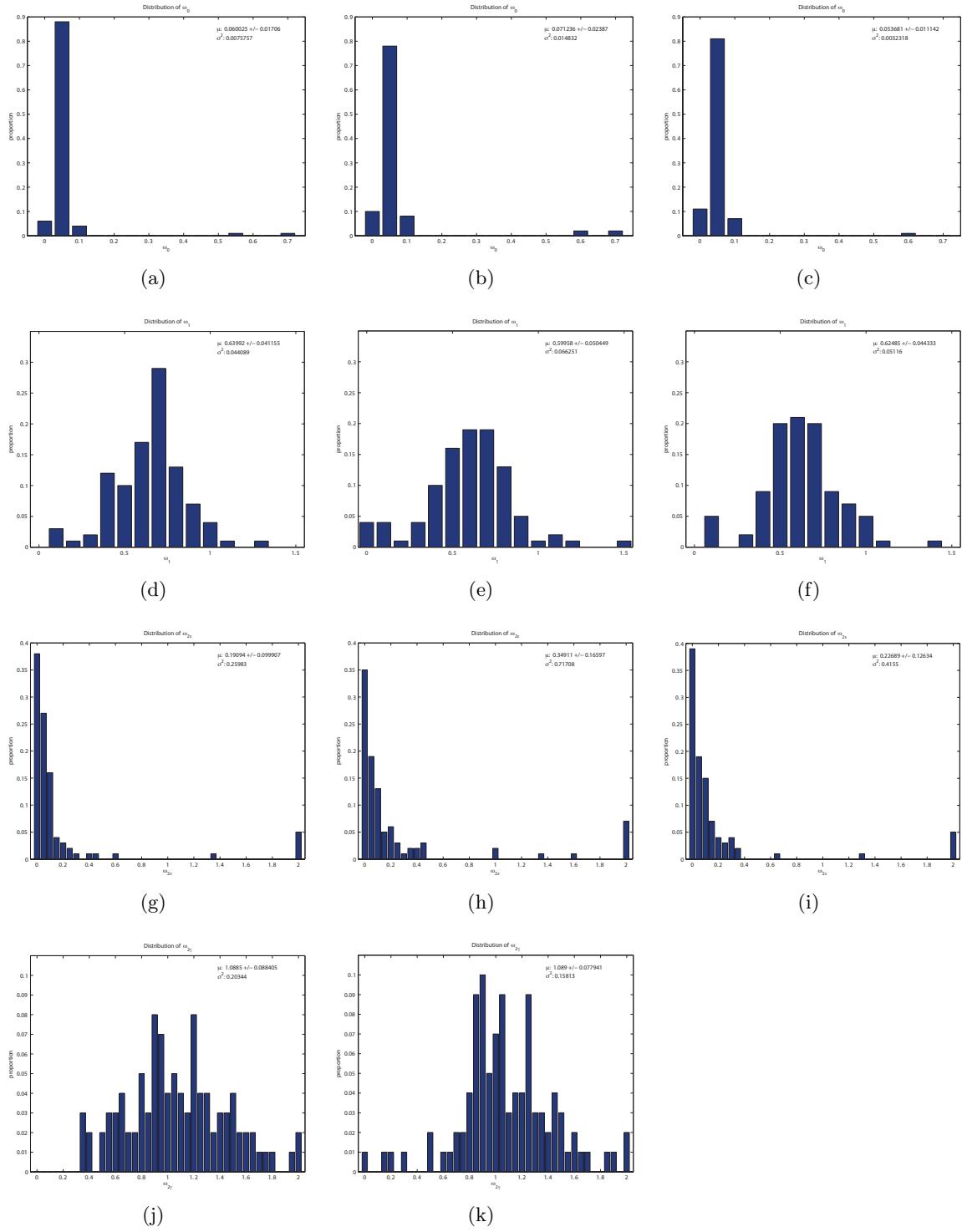


## Supplementary Material

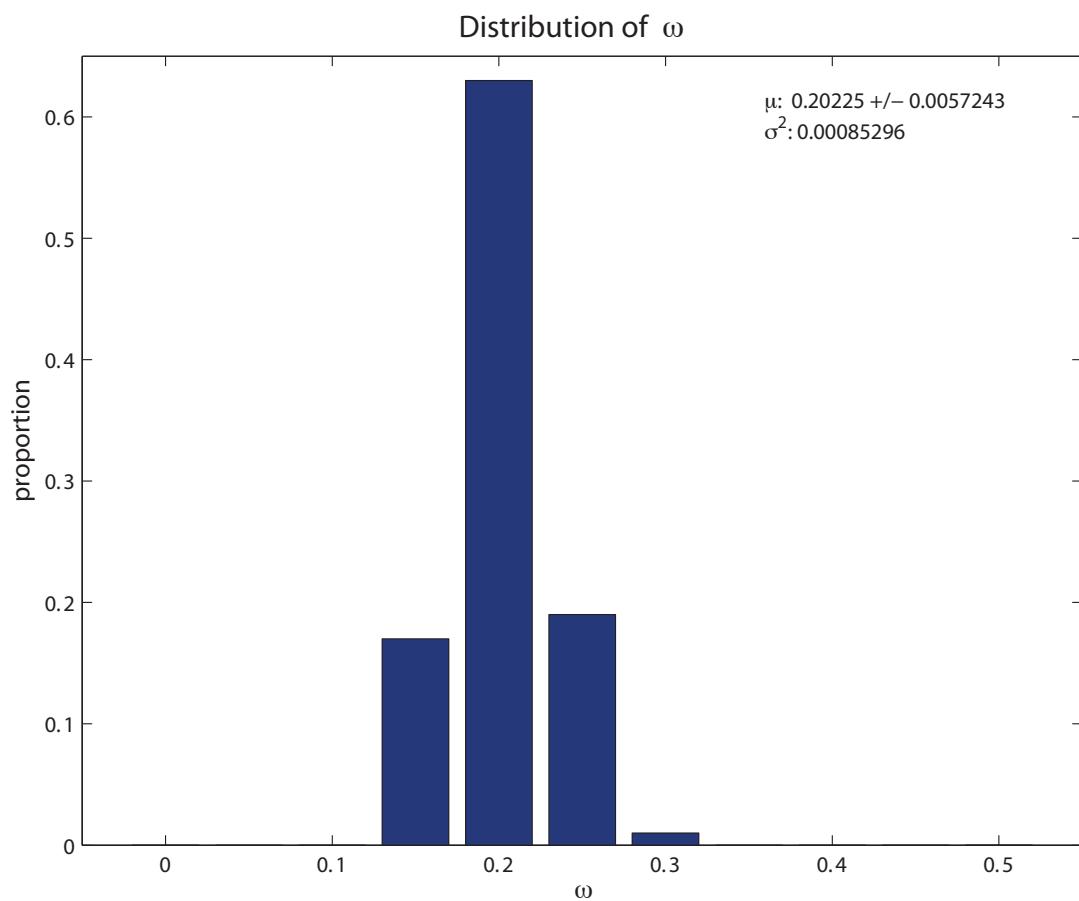
### Additional material for globin gene family example

Species	Gene	GenBank accession number
Alouatta seniculus	<i>HBE1</i>	L25367
	<i>HBG</i>	AF030097
Aotus azarai	<i>HBE1</i>	L25371
	<i>HBG</i>	U57044
Ateles geoffroyi	<i>HBE1</i>	L25368
Ateles paniscus	<i>HBG</i>	AF030093
Brachyteles arachnoides	<i>HBE1</i>	L25366
	<i>HBG</i>	AF030098
Callithrix jacchus	<i>HBE1</i>	L25363
	<i>HBG1/HBG2</i>	AF321384
Cebus apella	<i>HBG1</i>	U57043
Cebus olivaceus	<i>HBE1</i>	U18610
Cheirogaleus medius	<i>HBE1</i>	U11711
	<i>HGB</i>	M15758
Eulemur fulvus fulvus	<i>HBE1</i>	M15735
	<i>HBG</i>	M15757
Galago crassicaudatus	<i>HBE1</i>	M36304
	<i>HGB</i>	M36305
Homo sapiens	<i>HBE1</i>	U01317
	<i>HBG2</i>	U01317
Hylobates lar	<i>HBG1</i>	J05174
Hylobates syndactylus	<i>HBE1</i>	U64616
Lagothrix lagothrica	<i>HBE1</i>	L25358
	<i>HBG</i>	AF030094
Macaca mulatta	<i>HBE1</i>	M81364
	<i>HBG</i>	M19434
Otolemur crassicaudatus	<i>HBE1</i>	U60902
	<i>HBG</i>	U60902
Pan paniscus	<i>HBE1</i>	M81362
Pan troglodytes	<i>HBG2</i>	X03109
Pongo pygmaeus	<i>HBE1</i>	X05035
	<i>HBG1</i>	M16208
Saimiri sciureus	<i>HBE1</i>	L25354
Saimiri ustus	-	AF016984
Tarsius bancanus	<i>HBG</i>	AF0726810
Tarsius syrichta	<i>HBE1</i>	M81411

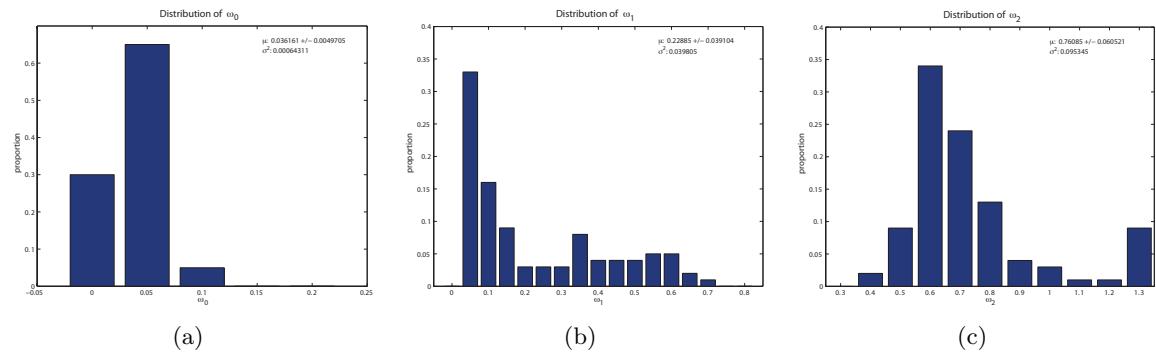
STable 1: List of genes used in the analysis of  $\epsilon/\gamma$ -globin family.



SFigure 1: The distribution of ML estimates for site classes of model MD ( $k = 3$ ) from three simulation runs with ALF under model MD for sequence length matching the real data (144 codons). (a)-(c):  $\omega_0$ ; (d)-(f):  $\omega_1$ ; (g)-(i):  $\omega_{2\epsilon}$ ; (j)-(k):  $\omega_{2\gamma}$  (histogram for run 1 included in manuscript)

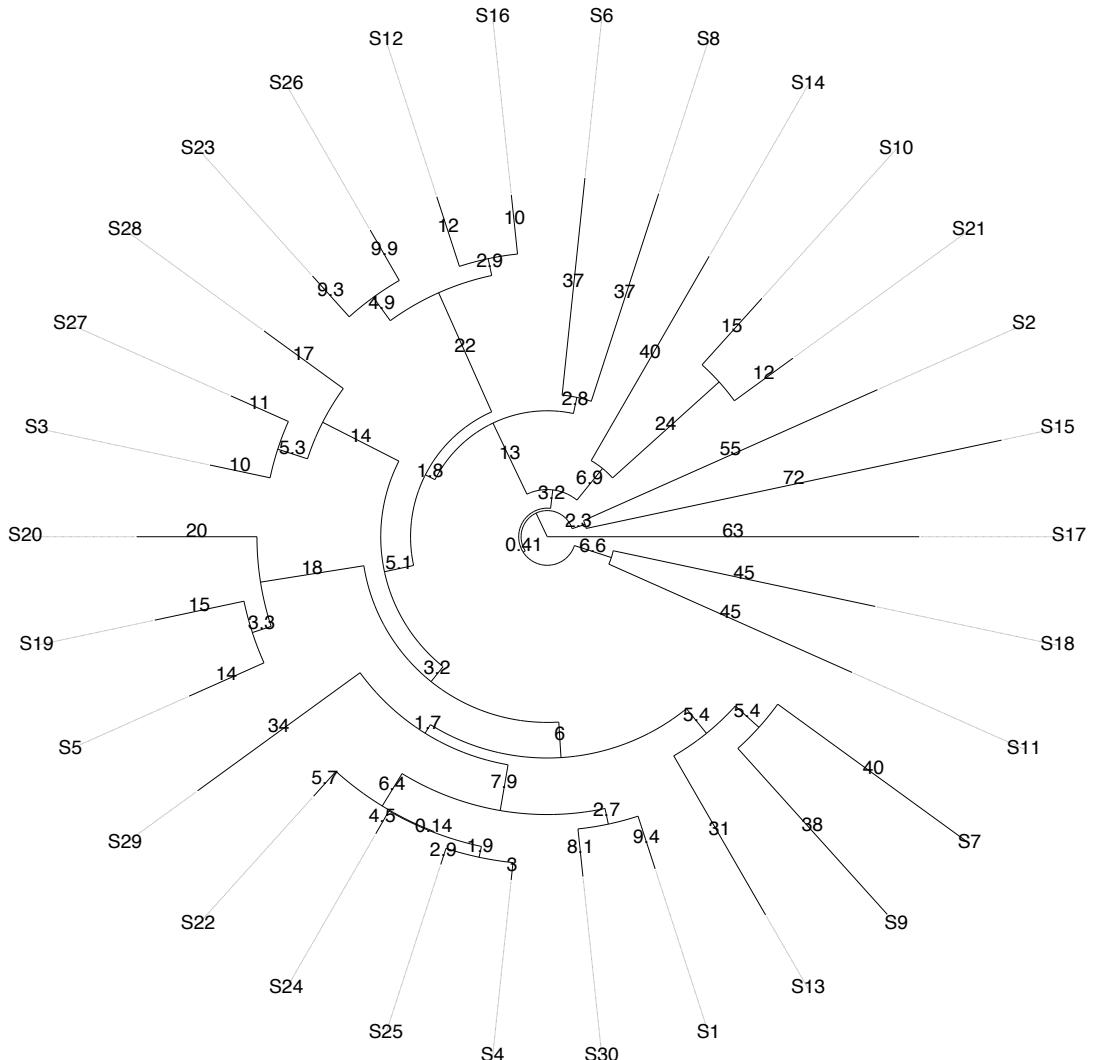


SFigure 2: The distribution of ML estimates for  $\omega$  under M0 from simulation with ALF under model MD for sequence length matching the real data (144 codons).



SFigure 3: The distribution of ML estimates for site classes of model M3 ( $k = 3$ ) from simulation with ALF under model MD for sequence length matching the real data (144 codons).

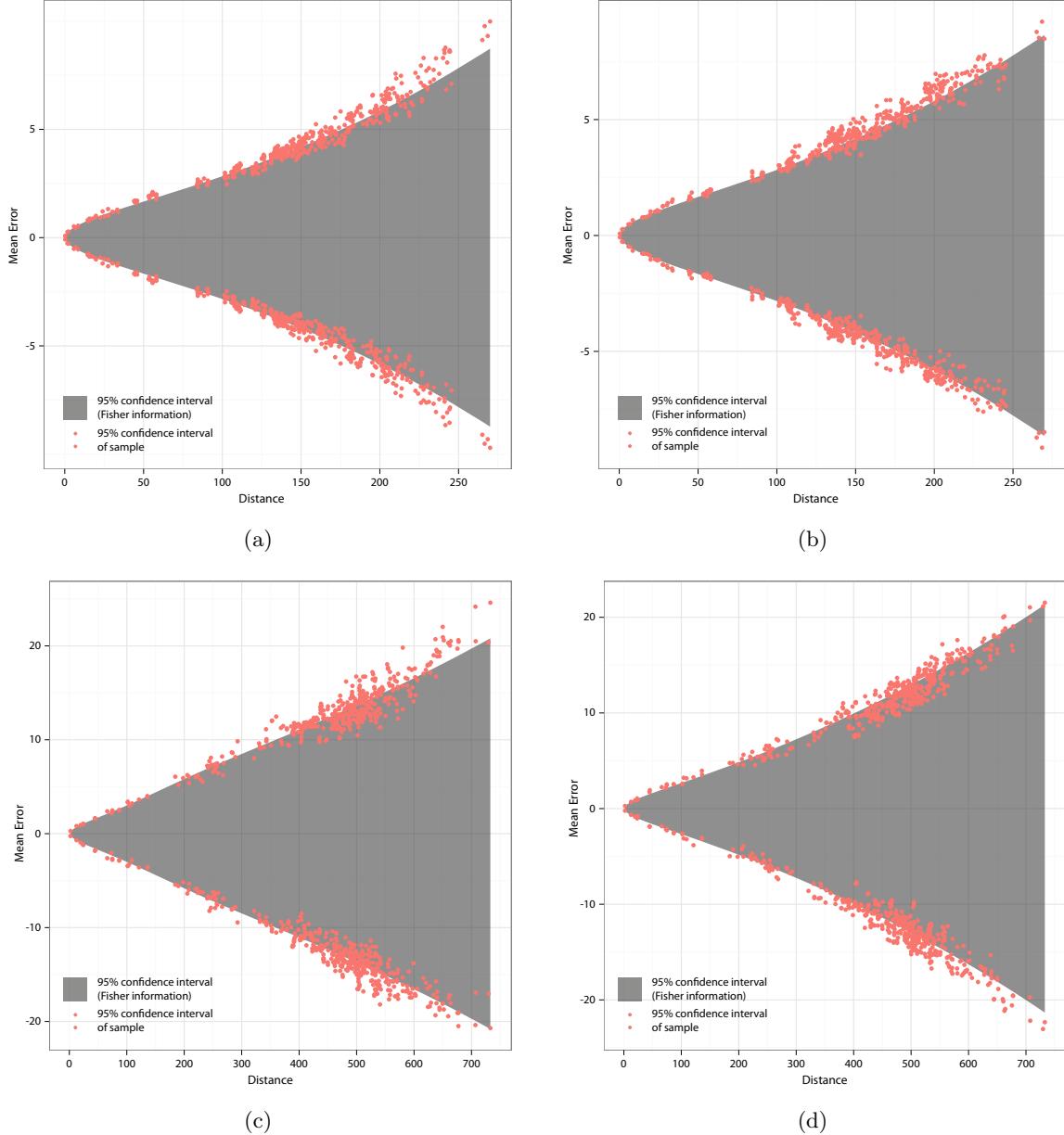
## Additional material for orthology inference with LGT



SFigure 4: Evolutionary scenario used for the analysis. We sampled 30 species from the tree of  $\gamma$ -proteobacteria.

## Validation

### Empirical models



SFigure 5: Validation of empirical amino acid and codon models. Variance of error of distance estimates ( $\hat{d} - d$ ) for pairwise distances of 30 simulated species. Based on 100 sequences of length  $N = 10000$  codons/amino acids per species. The grey area indicates the 95% confidence interval of the theoretical variance estimator based on the Fisher information. The red lines indicate the actual 95% confidence interval of the sample. (a) Gonnet matrix (GCB); (b) WAG matrix; (c) CodonPam; (d) ECM

## M-series models

Model	True Values	Replicate 1	Replicate 2
M0	$\kappa = 3.1 \quad \omega = 0.6$	$\kappa = 3.102 \quad \omega = 0.602$	$\kappa = 3.095 \quad \omega = 0.598$
M2	$\kappa = 3.1 \quad \omega_0 = 0$	$\kappa = 3.097 \quad \omega_0 = 0$	$\kappa = 3.111 \quad \omega_0 = 0$
	$\omega_1 = 1 \quad \omega_2 = 4.22$	$\omega_1 = 1 \quad \omega_2 = 4.161$	$\omega_1 = 1 \quad \omega_2 = 4.189$
	$p_0 = 0.404 \quad p_1 = 0.511$	$p_0 = 0.407 \quad p_1 = 0.508$	$p_0 = 0.403 \quad p_1 = 0.511$
M3	$\kappa = 3.1 \quad \omega_0 = 0.108$	$\kappa = 3.081 \quad \omega_0 = 0.108$	$\kappa = 3.123 \quad \omega_0 = 0.108$
	$\omega_1 = 1.211 \quad \omega_2 = 4.024$	$\omega_1 = 1.197 \quad \omega_2 = 3.823$	$\omega_1 = 1.222 \quad \omega_2 = 4.077$
	$p_0 = 0.604 \quad p_1 = 0.325$	$p_0 = 0.603 \quad p_1 = 0.323$	$p_0 = 0.604 \quad p_1 = 0.327$
M8	$\kappa = 3.1 \quad \omega = 3.385$	$\kappa = 3.108 \quad \omega = 3.172$	$\kappa = 3.087 \quad \omega = 3.314$
	$p = 0.222 \quad q = 0.312$	$p = 0.274 \quad q = 0.432$	$p = 0.272 \quad q = 0.423$
	$p_0 = 0.909$	$p_0 = 0.901$	$p_0 = 0.905$

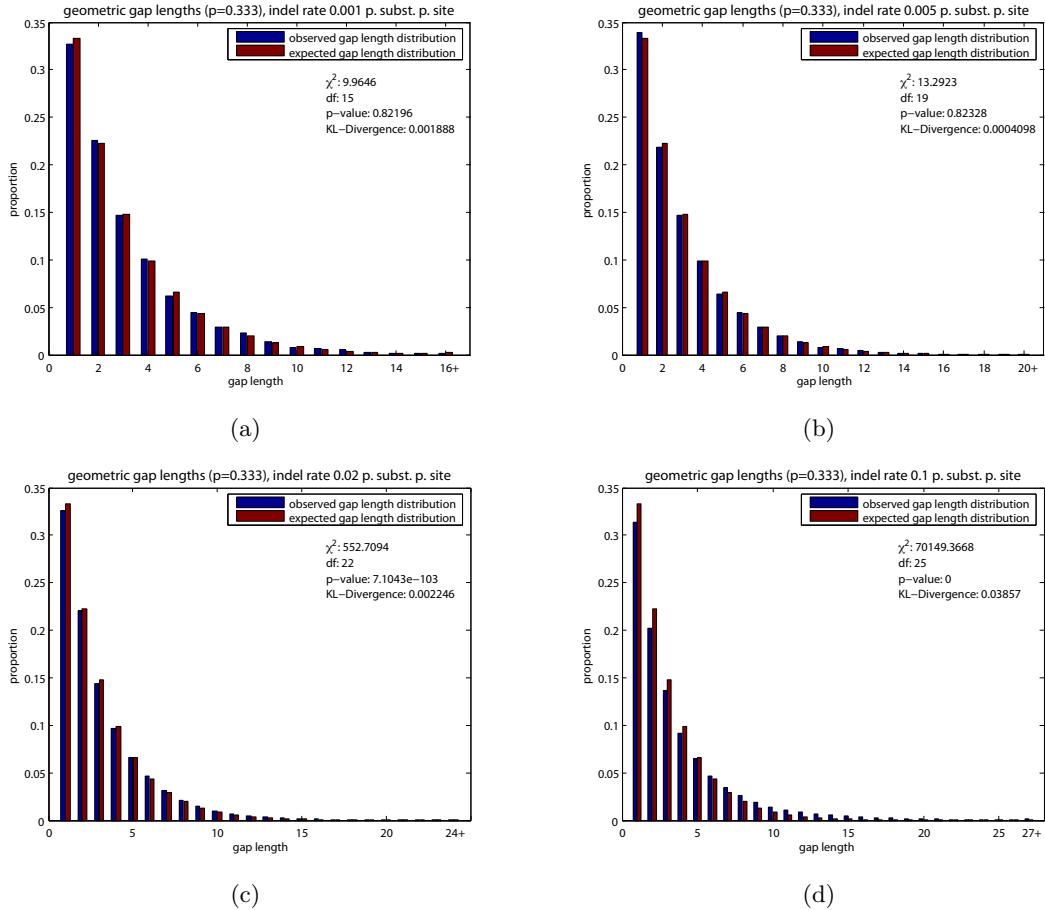
STable 2: Validation of M-series models. True parameter values and estimates from simulating two replicate datasets with 100000 codons.

## Nucleotide Models

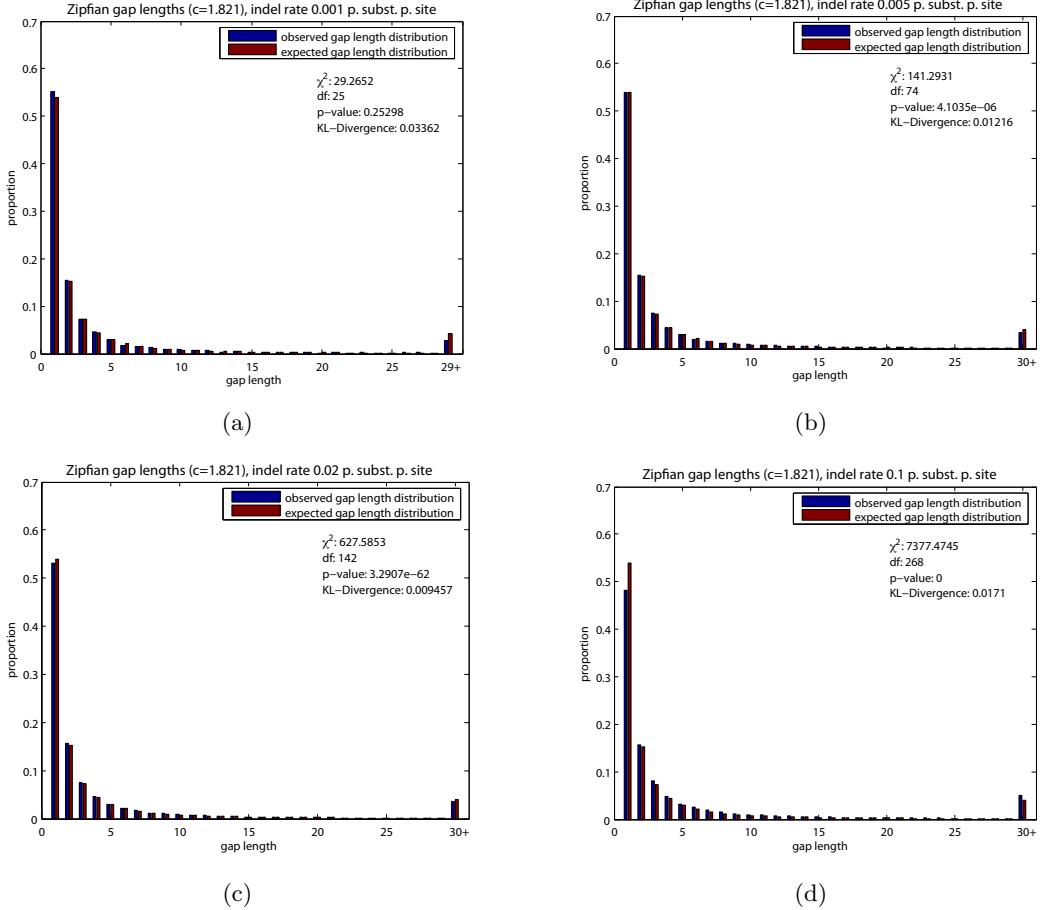
Model	True Values	Replicate 1	Replicate 2
F84	$tl = 7 \quad \kappa = 3.5$	$tl = 7.032 \quad \kappa = 3.470$	$tl = 7.125 \quad \kappa = 3.531$
	$\pi_T = 0.325 \quad \pi_C = 0.225$	$\pi_T = 0.324 \quad \pi_C = 0.225$	$\pi_T = 0.324 \quad \pi_C = 0.225$
	$\pi_A = 0.175 \quad \pi_G = 0.275$	$\pi_A = 0.175 \quad \pi_G = 0.276$	$\pi_A = 0.175 \quad \pi_G = 0.276$
GTR	$tl = 7 \quad a = 0.167$	$tl = 7.062 \quad a = 0.163$	$tl = 7.082 \quad a = 0.168$
	$b = 0.333 \quad c = 0.5$	$b = 0.331 \quad c = 0.495$	$b = 0.332 \quad c = 0.500$
	$d = 0.666 \quad e = 0.833$	$d = 0.650 \quad e = 0.830$	$d = 0.666 \quad e = 0.829$
	$\pi_T = 0.28 \quad \pi_C = 0.3$	$\pi_T = 0.280 \quad \pi_C = 0.300$	$\pi_T = 0.279 \quad \pi_C = 0.300$
	$\pi_A = 0.2 \quad \pi_G = 0.22$	$\pi_A = 0.200 \quad \pi_G = 0.220$	$\pi_A = 0.201 \quad \pi_G = 0.220$
HKY	$tl = 7 \quad \kappa = 3.125$	$tl = 7.039 \quad \kappa = 3.090$	$tl = 7.024 \quad \kappa = 3.119$
	$\pi_T = 0.28 \quad \pi_C = 0.3$	$\pi_T = 0.280 \quad \pi_C = 0.300$	$\pi_T = 0.280 \quad \pi_C = 0.300$
	$\pi_A = 0.2 \quad \pi_G = 0.22$	$\pi_A = 0.200 \quad \pi_G = 0.220$	$\pi_A = 0.200 \quad \pi_G = 0.220$
JC	$tl = 7$	$tl = 7.083$	$tl = 7.044$
	$\pi_T = 0.25 \quad \pi_C = 0.25$	$\pi_T = 0.250 \quad \pi_C = 0.250$	$\pi_T = 0.250 \quad \pi_C = 0.251$
	$\pi_A = 0.25 \quad \pi_G = 0.25$	$\pi_A = 0.251 \quad \pi_G = 0.249$	$\pi_A = 0.250 \quad \pi_G = 0.250$
TN93	$tl = 7$	$tl = 7.047$	$tl = 7.052$
	$\kappa_1 = 0.353 \quad \omega = 0.676$	$\kappa_1 = 0.349 \quad \kappa_2 = 0.675$	$\kappa_1 = 0.350 \quad \kappa_2 = 0.663$
	$\pi_T = 0.325 \quad \pi_C = 0.225$	$\pi_T = 0.325 \quad \pi_C = 0.226$	$\pi_T = 0.324 \quad \pi_C = 0.226$
	$\pi_A = 0.175 \quad \pi_G = 0.275$	$\pi_A = 0.175 \quad \pi_G = 0.274$	$\pi_A = 0.175 \quad \pi_G = 0.275$

STable 3: Validation of nucleotide models. True parameter values and estimates from simulating two replicate datasets with 100000 nucleotides.

## Gap length distributions

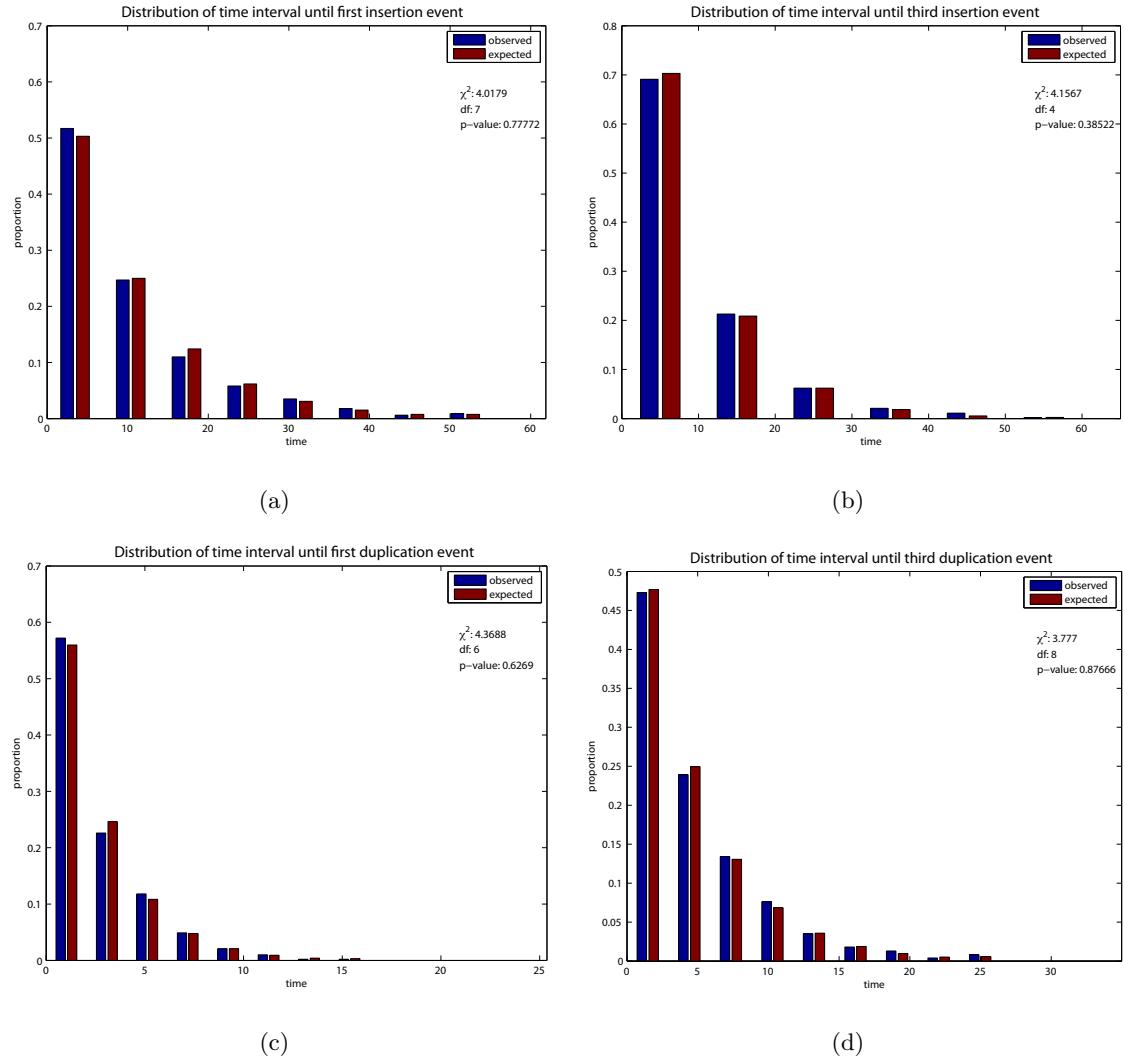


SFigure 6: Distribution of gap lengths for four evolutionary scenarios using geometrically distributed gap lengths ( $p = 0.333$ ) and differing indel rates (in indels per substitution per site). (a) 0.002; (b) 0.01; (c) 0.04; (d) 0.2. With increasing indel rates, longer gaps become more frequent than expected due to overlapping gaps whereas fewer short gaps are observed. This effect leads to a bias in the gap length distribution.



SFigure 7: Distribution of gap lengths for four evolutionary scenarios using Zipfian distributed gap lengths ( $c = 1.821$ ). Gaps were limited to a length  $< 500$  characters. Indel rates (in indels per substitution per site) were differed as follows: (a) 0.002; (b) 0.01; (c) 0.04; (d) 0.2. Due to the long tail of the Zipf distribution, the empirical distribution becomes skewed at lower indel rates compared to exponentially distributed gap lengths.

## Gillespie algorithm



SFigure 8: Distribution of waiting times between events for the Gillespie algorithm when simulating insertions and gene duplications. Observed distribution based on 1000 samples. (a) Waiting time to first insertion; (b) waiting time between second and third insertion; (c) waiting time to first gene duplication (on any of two branches); (d) waiting time between second and third consecutive gene duplication (on one particular branch).